NASA-CR-204200

7N-43-CR
030446

# A PERSPECTIVE ON THE ANALYSIS OF HYPERSPECTRAL DATA[1]*

David A. Landgrebe

School of Electrical Engineering
Purdue University
West Lafayette, IN 47907-1285, USA
Internet: landgreb@ecn.purdue.edu
Tel: (317) 494-3486; Fax: (317) 494-6440

## ABSTRACT

A very important recent technological development in sensor technology is the ability to construct multispectral imaging sensors with very large numbers (≈ 200) of spectral bands. A significant number of such hyperspectral sensor systems are now in use or under various stages of development in various countries. The existence of such sensors raises the question of how best to analyze the data with so many spectral bands.

Though existing conventional multispectral analysis methods will still be useful in the hyperspectral era ahead, it becomes clear that they fall short of the ideal in several senses. For example, conventional methods often prove computationally unreasonable for large numbers of bands, they may not enable the extraction of all the information from the data that might otherwise be possible, and they may prove excessively cumbersome for researchers in their work.

In this paper, after a brief introduction, some of the basic characteristics fundamental to the analysis process are outlined. This is followed by an example showing how they may be applied to hyperspectral data analysis.

## INTRODUCTION

Over the last three decades the field of land oriented optical remote sensor technology has seen the flight in space of at least two generations of multispectral imaging devices. Landsat 1 ushered in the modern era of multispectral sensing with the MSS, a four band sensor with its 80 meter pixels and a signal-to-noise ratio supporting a 6-bit gray scale. Landsat 4 and 5 carry Thematic Mapper, a second generation sensor with its seven bands of 30 meter pixels more broadly distributed over the optical region and a signal-to-noise ratio supporting 8-bit data. With some modest enhancements, Landsat 6 and 7 will carry similar capabilities.

Advances in the solid state devices field since the mid-seventies when Thematic Mapper was designed, have made possible significantly more advanced sensors. A considerable number of multispectral devices with as many as 200 spectral bands are now flying in aircraft or under various stages of design or construction in a number of countries. Because of the large number of bands, these devices are now referred to as hyperspectral sensors, and in spite of the much narrower spectral windows being used, they can provide signal-to-noise ratios supporting 10-12 bit data systems with spatial resolutions of a few 10's of meters IFOV from orbit.

Such a large jump in data complexity requires a renewed focus on data analysis technology, for though existing approaches can still be used with such data, if the full information delivery potential of such sensors is to be realized, it is reasonable to speculate that paralleling advances in analysis methods are needed. Thus we turn to a re-look at the fundamentals of multivariant data analysis.

## DATA ANALYSIS PRINCIPLES

As a result of fundamental engineering research over the years, much has been learned about the process of analysis of complex data. Results drawn from the fields of the communication sciences, pattern analysis, and signal processing are particularly relevant to the remote sensing problem. It is possible to put forth some basic principles useful as a point of departure in addressing the unprecedented complexity of the new data.

**Data Representation.** Perhaps foremost of these principles is that the analysis perspective must begin with a rigorously defined but broadly applicable means for mathematically representing the data. The mathematical means for representing the data must be such that it does not ignore any aspect of the data that might be information-bearing. A number of different data representations have been in common use in the field, depending upon the application problem and the background of the analyst. Straightforward schemes that view multispectral data strictly in terms of spectra, i.e., a graph of response values verses the band number, have been found to be very practical for many problems.

However, it is becoming increasingly apparent that there is significant useful information contained not only in the spectral variations themselves, but how the signals co-vary from band to band, and such information can easily be overlooked in viewing data as simple deterministic spectral responses. For example, the second order variations of spectral responses have been shown to be of increasing importance as the dimensionality of the data increases [1,2]. Thus, to preserve the ability to express these variations, the use of a mapping from "spectral space" to a finite dimensional "feature space" is suggested. Such a representation has found some use in past years [3,4] but it will become even more important with the arrival of the more complex data provided by hyperspectral sensors. The concept is simple, as illustrated in Figure 1, but quite general, effectively mapping a continuous function into a discrete finite dimensional space in a bilateral, lossless fashion.

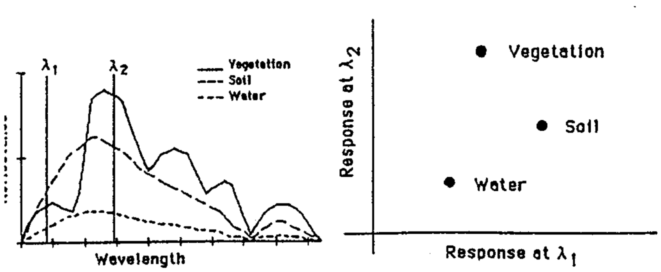


*Figure 1. Three example spectral responses mapped to a multivariant feature space. Though sampling at only two wavelengths is shown here, sampling at any number of wavelengths could be used, resulting in a correspondingly higher dimensional multivariant space.*

Next we note that though it is perhaps convenient to think of a given material as having a single spectral response, as implied by the spectral curves of Figure 1, in reality, pixels are mixtures

1 This work was supported in part by NASA under Grant NAGW-925.

*Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS'93), Tokyo, pp 1362-4, August 1993.

of a number of constituents under slightly different illumination and observation circumstances. Thus, when observed at a high level of precision and detail, a given material provides not a single spectral response but one of a family of responses. It is this family that is diagnostic of the material, rather than any one of its members alone.

**Characteristics of Analysis Algorithms.** Next one must consider what are the significant characteristics of analysis algorithms. Here some key fundamental principles come to play. It is a fundamental principle of measurement technology that relative measurements can be more precisely made than can absolute ones. For example, the time interval between two epochs can be measured to parts in $10^{12}$ or better, while the absolute time (of day) can be measured to parts in $10^3$. The same principle is involved in analysis of multispectral data. It is easier to decide between two alternative possible materials than it is to identify a material on an absolute basis.

This suggests that analysis algorithms that discriminate between alternatives should be more powerful than those which identify a material outright. The implication arising from this principle is that, to gain full advantage of the data, one must construct an exhaustive list of possible outcomes to the particular analysis, and that relative pattern classifiers should be more effective than absolute ones.

In a practical sense the analysis process is in effect the merging of prior knowledge with the data to be analyzed. This prior knowledge may be of a subjective nature, such as the information one possesses as a result of being an expert in the particular sub-discipline of Earth science involved, or of having specific detailed knowledge of the ground area from which the data were gathered. It may also consist of quantitative data that has previously been gathered about the particular scene or the particular materials of interest. It is also fundamentally true that the more of either of these kinds of information that can be brought to bear upon the analysis task, the greater the information that can be derived from the data.

Further, especially regarding prior knowledge of a quantitative nature, one must somehow reconcile the circumstances of data collection of the prior information with those of the current data. Perhaps the most obvious way of doing this is to calibrate both data sets, however, this has proven to be a daunting task and is itself subject to inaccuracies. As the complexity of data increases, the importance of accurate calibration increases, but so does the difficulty of achieving the increased accuracy that the increased level of detail requires. Thus, if other methods for reconciling prior and current data can be used, or if analysis methods can be found which are less sensitive to noise including the inaccuracies of calibration, they should be advantageous.

Finally, regarding analysis algorithm characteristics, the circumstances of the remote sensing situation require that, if possible, it is desirable that the analysis method not require concomitant data collection from the ground. This is a desirable characteristic, because it is necessary that an analyst be able to analyze data sets from any part of the world. Note that there is not a requirement for the analysis to be automatic in the sense of no human participation in the analysis process. Indeed, as pointed out above, the greater the prior knowledge, be it quantitative or be it subjective resulting from human expertise, the greater the possible performance of the analysis process.

Summarizing then,

- Materials to be identified should be effectively modeled, and in a manner which associates classes of materials of interest with families of spectral responses (rather than individual spectra).
- A relative scheme is preferred over an absolute one.
- Some means of reconciling current data with prior data is required.
- An algorithm that is quantitatively oriented but which provides for the use of human expertise in an objective and efficient manor is highly desirable.

### But What About Hyperspectral Data

It has been indicated that the emergence of hyperspectral sensors is really driven by a technological development. One question often raised is, "Is there a demonstrated 'requirement' for hyperspectral data?" Though it is frequently assumed that requirements should precede technological capabilities, it is probably true that for most of the greatest scientific advancements, it happens the other way around. Even so, is there a demonstrated need for hyperspectral data given it has become possible?

No doubt the ice-breaker in this regard has been imaging spectroscopy [5]. Simply stated, this is an attempt to do from space what the chemical spectroscopist does in the laboratory, i.e., identify materials by looking for narrow absorption lines in the spectra that are diagnostic of specific molecules. It is an important technique, not only because it works for certain applications, but because the method can be quickly grasped and clearly understood by those who have only peripheral knowledge of remote sensing. Thus it has been successful in launching a broader interest in this advancement of technology.

The question remains, however, as to what the ultimate potential of the technology is and how to approach it. It is reasonable to hypothesize that with the order of magnitude increase in both the number of bands and in the signal-to-noise ratio, one should be able to increase the number of materials, the level of detail of materials, or the identification accuracy, or perhaps all three of these. But how does one deal most effectively with 200 dimensional data?

The answer to this question is a matter for research for which there is no single answer at this time. One approach to the problem is illustrated in Figure 2.

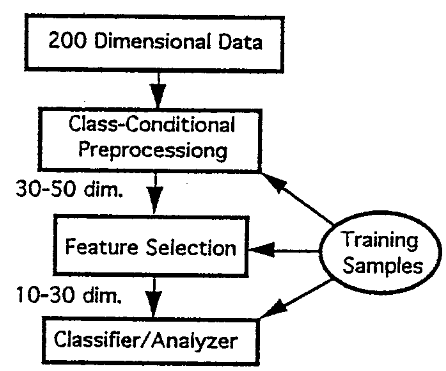


*Figure 2. A paradigm for using 200 dimensional data effectively. This analysis logic uses a case-specific pre-processing step that reduces the dimensionality to a range that can be dealt with effectively via conventional feature selection.*

The concept in this approach is to use a preprocessing step utilizing class conditional information such that the dimensionality is reduced without loss of relevant information.

To illustrate this technique we will briefly outline the method and results of analysis of an AVIRIS [6] data set. The problem is to produce a geologic map of the Cuprite mining district in southwestern Nevada. This area has been overflown several times by AVIRIS, and has been previously mapped on several occasions. The intent this time is to use methods applying the above fundamentals to illustrate the ability to function effectively in the face of noise. Therefore, a data set was chosen which was collected early in the life of AVIRIS (1987) when the signal-to-noise ratio was much lower than it is today. Further, the classification, itself, was done without any adjustment for atmospheric effects or any other calibration or preprocessing of the data.

In this analysis, four minerals were of primary interest: Alunite, Buddingtonite, Kaolinite, and Quartz. The analysis was done using the MultiSpec[1] software system together with Matlab[2] implemented upon a Macintosh[3] computer, all of which are publicly available. The technique used was to select a number of training samples for each of these minerals using interpretive information as might be possess by an experienced geologist. This was done with the aid of a Matlab-implemented log residue transform capability to examine candidate training pixels that contain the characteristic absorption spectra for each of the four minerals. The log-residue adjustment is used to adjust the shape of the radiance spectra to be more similar to reflectance spectra, as might be obtained in laboratory measurements, thus increasing the manual interpretability of the spectral. Note that this method of developing training for a pattern recognition algorithm allows for the effective use of expert knowledge of the geologist/analyst, but in a manner that does not compromise the quantitative nature of the data. Note also that the use of training samples inherently reconciles the prior knowledge with the observation conditions of the particular data set.

Recall that best performance should occur when a relative classification scheme is used. This implies the need for an exhaustive list of classes. Using the training set developed to this point for the four desired classes, a preliminary classification was carried out. Besides producing a Classification Map, one of the capabilities of MultiSpec is to produce a Likelihood Map, i.e., a display in which each pixel is exhibited in a color shade corresponding to the degree of membership which that pixel has to the (maximum likelihood) class to which it has been assigned. All pixels having a low likelihood of class membership suggest that they may be members of other classes not yet defined. Using this display, again combined with the analyst knowledge of geology, additional training the remaining classes of tuff, alluvia and playa were defined.

While the absorption features used in selecting the training data for the four minerals were useful in finding class training data, they are not necessarily the best for discrimination between the classes of the scene. To find the most effective features, the decision boundary feature extraction method [7,8] was used. Finally, the ECHO (Extraction and Classification of Homogeneous Objects) classifier of MultiSpec was used to obtain the final result. The ECHO classifier is a spectral/spatial classifier which first partitions the scene into statistically similar groups of adjacent pixels, called objects. The objects are then classified by a maximum likelihood sample scheme. The final result is shown in Figure 3 (original in color).

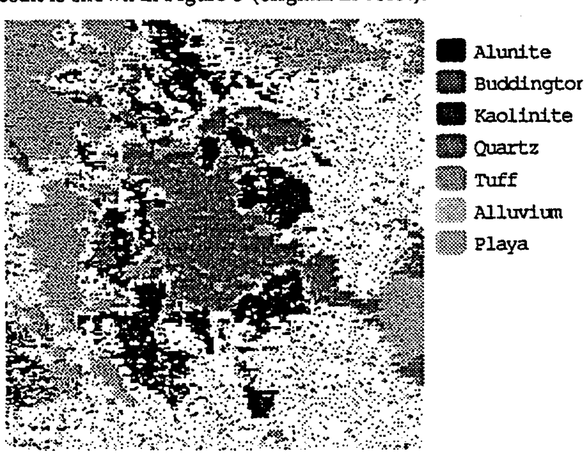


*Figure 3. ECHO classification of Cuprite area. (Original in color)*

This result compares favorably on a subjective basis with other mappings of the area done with more extensive procedures and high signal-to-noise ratio data. Without detailed ground truth from the area, a quantitative assessment of accuracy is difficult to obtain directly. Two indirect methods commonly use in the field of pattern recognition are the re-substitution method (known to be optimistically biased) and the leave one out method (known to be pessimistically biased). By these methods, the classification accuracy is measured as 93% and 92% respectively.

In summary, this example implements most or all the basic principles described earlier and for the trouble, it achieves results comparable to other methods in the face of greater levels of noise and without using several of the computationally intensive data adjustment methods commonly seen as required. Given the high accuracy of the result even without any preprocessing adjustments to the data, this suggests that hyperspectral data may have greater potential than is shown by this particular example. The problem, itself, is not as challenging as might be desired for a more thorough test, and one might hope that a problem with more detailed classes might be undertaken with a reasonable chance for success. This, together with further improvements in the algorithms used, remain for the future.

## References


[1] Chulhee Lee and David A. Landgrebe, "Analyzing High Dimensional Data," International Geoscience and Remote Sensing Symposium (IGARSS'92), Houston, TX, May 26-29, 1992.

[2] David A. Landgrebe, "On the Use of Stochastic Process-Based Methods for the Analysis of Hyperspectral Data," International Geoscience and Remote Sensing Symposium (IGARSS'92), Houston, TX, May 26-29, 1992.

[3] Swain, P.H., and Davis, S.M. (editors), *Remote Sensing: The Quantitative Approach* McGraw-Hill, 1978, Chapter 1.

[4] Richards, John A., *Remote Sensing Digital Image Analysis, An Introduction* Springer-Verlag, 1986, Chapter 3.

[5] Goetz, A.F.H., Gregg Vane, Jerry E. Solomon, and Barrett N. Rock, Imaging Spectrometry for Earth Remote Sensing, Science, Vol. 228, pp1147-1153, June, 7 1985.

[6] Chrien, T.G., M.L. Eastwood, C.M. Sarture, R.O. Green, and W. M. Porter, Current Instrument Status of the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS), Proceedings of the Third Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Workshop, May 20-21, 1991, NASA/JPL, May-91.

[7] Lee, Chulhee; Landgrebe, David; "Feature Extraction Based on Decision Boundaries," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 3, pp 388-400, April 1993.

[8] Lee, Chulhee and David A. Landgrebe, "Feature Selection Based On Decision Boundaries," International Geoscience and Remote Sensing Symposium, Espoo, Finland, June1991.


---

1 MultiSpec (© Purdue Research Foundation) is a multispectral data analysis program available from the author.

2 Matlab is a trademark of the MathWorks, Inc.

3 Macintosh is a trademark of the Apple Computer Corporation.